UIUC/UTD: Weakly Supervised Cause Analysis via Semantic Lexicon Construction

Muhammad Abedin, Vincent Ng, Latifur Khan

Agenda

- Problem Definition
- Existing Research
- New Approaches Explored
- Ongoing and Future Works

Problem Definition

- Given an Air Traffic Incident Report, our task is to identify which Shaping Factor(s) contributed to the occurrence of the incident.
- Input:
  - Air Traffic incident report
- Output:
  - Factors that led to the incident

Domain: ASRS

- We are working on Air Traffic Incident Reports from the Aviation Safety Reporting System (ASRS)
  - Each time an air traffic incident (like altitude deviation, runway incursion etc.) occurs, the pilot has to file a report on the incident
    - Each report contains a plain text narrative describing the incident, and various other data
  - Publicly available at asrs.arc.nasa.gov

Shaping factors

- [Posse 2005] identifies 14 Shaping Factors that may be the cause of the incidents.
- These are:
  - Attitude, Communication Environment, Duty Cycle, Familiarity, Illusion, Physical Environment, Physical Factors, Preoccupation, Pressure, Proficiency, Resource Deficiency, Taskload, Unexpected, Other, Attitude, Communication Environment, Duty Cycle, Familiarity, Illusion, Physical Environment, Physical Factors, Preoccupation, Pressure, Proficiency, Resource Deficiency, Taskload, Unexpected and Other.

Attitude	Any indication of unprofessional or antagonistic attitude
Communication Environment	Interferences with communications in the cockpit
Duty Cycle	A strong indication of an unusual working period
Familiarity	Any indication of a lack of factual knowledge
Illusion	Anything leading to erroneous perception
Physical Environment	Unusual physical conditions that could impair flying or make things difficult
Physical Factors	Pilot ailment that could impair flying or make things more difficult
Preoccupation	A preoccupation, distraction, or division of attention that creates a deficit in performance
Pressure	Psychological pressure
Proficiency	A general deficit in capabilities
Resource Deficiency	Absence, insufficient number, or poor quality of a resource
Taskload	Indicators of a heavy workload or many tasks at once
Unexpected	Something sudden and surprising that is not expected

Existing Research

- **[Posse 2005]** describes text normalization and pattern extraction methods to recognize the shaping factors
- **[Ferryman 2006]** applies the method in [Posse 2005] to a set of reports and uses the results to identify the shaping factors responsible for different anomalies

Posse 2005

- Brainstorming session:
  - Involved experts on aviation safety, human factors, English language and data analysis
  - For each shaping factor:
    - Identified seed keywords, simple expressions and template expressions.
- Manually encoded the expressions and template patterns as JAPE rules
- Given a report, identified responsible shaping factor(s) by matching patterns
  - Experimented on 20 reports

Ferryman 2006

- Applied the method in [Posse 2005] to 17155 reports to identify the shaping factors
- From the incidents described in the reports and the shaping factors identified in the previous step, computed which shaping factors are most associated with which incidents

Challenges

- Can we construct expressions or features identifying the shaping factors without extensive involvement of domain experts and their knowledge?
- The goal of our research is to reduce the dependency on domain specific knowledge by learning expressions only from the text and the example keywords for the shaping factors, utilizing natural language features like part-of-speech and sentence structure.

Seed Keywords List

- From the narratives of 233 reports, , we have chosen 243 seed words for the 14 shaping factors.
- Additionally, we have included the categories People and Equipments as control categories, adding 11 more keywords.

Sample Seed Words

- For example, seed words for shaping factor ***Proficiency*** are:
  - mistakes, mistake, mistaken, error, new hire, inexperience, wrong course, relative inexperience, first YR, mistook, forgotten, misinterpreted, less than 100 hours, newly rated, training, recent pilot, inadvertently, bad turn, train, new Captain, MISINTERPED

Motivation

- How do we guess the meaning of an unknown word?
  - We try to guess the meaning from the context
  - If we observe that an unknown word X is frequently being used in a context similar to another known word Y, we would guess that X has a meaning similar to Y
- Contexts can be captured with extraction patterns.

Preprocessing

- The ASRS report narratives are written in an informal manner, containing numerous abbreviations, acronyms and jargon
  - The text needs to be normalized by expanding the abbreviations before the it can be analysed properly
- Also, the text is in all upper case
  - The performance of part of speech tagging depends in some cases on the proper case of the words

Case Sensitivity Example

- All upper-case text (incorrect tagging)
  - WE/PRP WERE/VBD IN/IN **OUR/NNP** CLIMB/NNP **FROM/NNP** CMH/NNP **TO/NNP** DFW/NNP ;/: **WITH/VB** A/DT CLEARANCE/NNP **TO/NNP** FL310/NNP ./.
- Partially case restored text (correct tagging)
  - we/PRP were/VBD in/IN our/PRP$ climb/NN from/IN CMH/NNP to/TO DFW/NNP ;/: with/IN a/DT clearance/NN to/TO FL310/NNP ./.
- 5 mis-tagging in this sentence!

Partial Case Restoration

- For our experiments, we have performed a partial case restoration as follows:
  - For each word W in the text:
    - Look up W in the acronym expansion table. This tables contains the expansions in correct case, so if W is found, it is expanded to correct case.
    - Look up W in English language dictionary. If it is found, it converted to lower case.
    - If W is neither in the expansion table nor in the dictionary, it is simply left upper case.

Lexicon Building Process

Step – 1: Identify Nouns and Adjectives in the text

Step – 2: Identify Patterns that extract the Nouns and Adjectives

Step – 3: Using the Patterns as features, classify the words into the different Shaping Factors

Lexicon Growing

- Using the extraction patterns, we build the shaping factors lexicon using the Basilisk-MCAT framework described in [Thelen 2002].
- Basilisk-MCAT is a semantic lexicon learning framework that automatically learns new keywords for a set of classes.
  – It requires a starting set of seed words for each class.

Basilisk-MCAT

- Input:
  - Words from the text
  - Patterns that extract the words in the text
  - Classes with seed words for each class
- Output:
  - For each class, a specific number of new words related the to seed words for that class
- Run multiple iterations, growing the lexicon gradually.

Step-1: Construct Pattern Pool

- Construct a pattern pool consisting of the patterns that are most likely to extract words in the given classes.

```
global_pattern_pool=empty
For each class C
    Compute R = (F / N) * log(F) for each
      pattern [N=number of times that
      pattern P extracts a word, F=number
      of times that pattern P extracts a
      word in class C]
    Sort patterns by R in descending order
    Put the top (20 + iteration number)
      patterns into global_pattern_pool
```

Patterns For Attitude

Pattern	RlogF
lulled into <X>	3.6
created some <X>	2.78
was some <X>	2.74
<X> and misunderstanding	2.63
caused some <X>	2.38
some initial <X>	2.32
guard against <X>	2.03
to eliminate <X>	1.87
cone of <X>	1.76
<X> over whether	1.62
Diode to <X>	1.44
leads to <X>	1.42
been some <X>	1.41
lead to <X>	1.37
moment of <X>	1.36
eliminate any <X>	1.34
be some <X>	1.32
<X> over where	1.29
<X> over who	1.29

Step – 2: Construct Word Pool

- Construct the word pool by combining the words that are extracted by the patterns in the pattern pool

```
global_word_pool=empty
For each pattern P in
  global_pattern_pool
    For each word W extracted by P
      If W is not assigned to any class
        Add W to global_word_pool
```

Word Pool Example

Pattern	Word
created some <X>	confusion
created some <X>	stress
created some <X>	problems
created some <X>	complacency
created some <X>	flooding
lulled into <X>	security
lulled into <X>	landing
lulled into <X>	thinking
lulled into <X>	complacency
lulled into <X>	sleep
lulled into <X>	indifference
lulled into <X>	shortcuts
lulled into <X>	'this
was some <X>	part
was some <X>	time
was some <X>	cause
was some <X>	failure
was some <X>	light
was some <X>	distraction

Step – 3: Compute Word Scores

- For each word in the word pool, compute the probabilities of its being in each category

```
For each word W in
  global_word_pool
    For each Category C
      Compute AVGLOG(W, C) = SUM(log2(Fj +
        1))/P
          P = number of patterns that extract W
          FJ = number of words in category C
            that pattern J extracts
```

Scores For The Word “flooding”

Category	AvgLog
Physical Environment	6.51
Equipment	6.41
Communication Environment	5.49
People	5.29
Physical Factors	5.09
Resource Deficiency	4.86
Attitude	4.7
Preoccupation	4.52
Duty Cycle	4.46
Familiarity	3.81
Unexpected	3.58

Step – 4: Assign Classes

- For each word in the word pool, the class for which the probability is highest is assigned to the word

```
For each word W in
  global_word_pool
    Assign C as the class of W for which
      AVGLOG(W, C) is maximum
```

- For example, in the previous example "flooding" will be assigned the class "Physical Environment"

Step – 5: Add Seed Words

- For each class, add the five words with the highest probability of belonging to that class to that class's seed word set.

```
For each class C
    Add five words with the highest
      values of AVGLOG(W, C) to the seed
      words set of class C
```

- Go back to step – 1.

Analysis

- Not too many words were added, nor was the accuracy encouraging
- We decided to explore a different method of classifying the words:
  - Probabilistic classification
  - For each word *W* in the word pool, we calculate the probability that it belongs to shaping factor *S* for each shaping factor *S*, and assign *W* to the *S* for which the probability is highest.

Problem With AvgLog Metric

- The AvgLog metric considers the absolute number of category words extracted by the patterns that extract the word.
- However, it does not consider frequency of extraction.
  - For example, if pattern P occurs 1000 times, but extracts words in category C only 5 times, it is unlikely that P is related to C.
  - Similarly, if word W occurs 1000 times, but is extracted by pattern P only 5 times, P should have a small influence on classification of W.

Example

- Consider a word W extracted by patterns P_1, P_2 and P_3. If each of them extract 5 words in category C, AvgLog=2.32. However, the patterns extract words in C only a small fraction of their occurrence.

Patterns that extract W:	P_1	P_2	P_3	
Number of category words extracted by the pattern P_i:	5	5	5	
Log2(F+1):	2.32	2.32	2.32	AvgLog= 2.32
Number of times that W is extracted by the pattern P_i:	10	20	70	Total= 100
Number of times pattern P_i occurs in the text:	100	500	1000	
Number of times a word in category C is extracted by the pattern:	5	5	5	

Probabilistic Classification

- Let W be a word, P be a pattern that extracts W in the text and C be a shaping factor. Then we have:
    - Pr(P|W)
        - Probability that P extracts W in text
    - Pr(C|P)
        - Probability that P extracts a word related to C
    - $Pr(C|W) = SUM_{\text{All P that extract W}} [Pr(C|P) * Pr(P|W)]$
        - The probability that W is related to C
  - The Shaping Factor C for which Pr(C|W) is maximum is the class for W.

Probability Vs AvgLog

- This method does not suffer from the problem since it depends on the probability of the word's being extracted by the patterns and the patterns' probability of extracting words in the category.

Patterns that extract W:	P_1	P_2	P_3	
Number of category words extracted by the pattern P_i:	5	5	5	
Log2(F+1):	2.32	2.32	2.32	AvgLog= 2.32
Number of times that W is extracted by the pattern P_i:	10	20	70	Total= 100
Number of times pattern P_i occurs in the text:	100	500	1000	
Number of times a word in category C is extracted by the pattern:	5	5	5	
$Pr(P_i\|W)$:	0.1	0.2	0.7	
$Pr(C\|P_i)$:	0.05	0.01	0.01	
Pr(C|W):	0.01	0.0020	0.0035	SUM(Pr(C|W))= 0.01

Patterns Explored

- We have experimented with two types of patterns:
  - N-gram based patterns, and
  - Parse tree based patterns.

N-gram Based Patterns

- Given a sentence, a N-gram is any sentence fragment that has N words.
- For each noun and adjective that appear in the corpus, we use two N-gram patterns:
  - The preceding N words + <X>, and
  - <X> + the succeeding N words.
  - For example, in the sentence ”... a *solid line of thunderstorms was detected ...*” the 2-gram patterns for ”thunderstorms” would be:
    - line of <X>
    - <X> was detected

N-gram Pattern Example

- Sentence:
  - ”approaching the **ATL area** a **solid line** of **thunderstorms** was detected in the **vicinity** of the **airport**.”
- Words (nouns and adjectives) and 2-gram patterns:
  - **ATL**: approaching the <X>, <X> area a
  - **area**: the ATL <X>, <X> a solid
  - **solid**: area a <X>, <X> line of
  - **line**: a solid <X>, <X> of thunderstorms
  - **thunderstorms**: line of <X>, <X> was detected
  - **vicinity**: in the <X>, <X> of the
  - **airport**: of the <X>

Extracting Phrases

- The examples of the shaping factors contain phrases like *long day*, *last leg* etc.
- Thus, extracting only noun and adjective words is not sufficient, we must extend our work to extract noun phrases and adjective phrases as well.
- The N-gram patterns can easily be extended to extract phrases.

N-grams Patterns For Phrases

- We use the CRFChunker tool to identify the phrases in the corpus.
  - We removed any articles (a, an, the) and possessive pronouns and adjectives (my, his, etc.) from the beginning of the phrases.
- For each noun phrase and adjective phrase that appear in the corpus, we use two N-gram patterns:
  - The preceding N words + <X>, and
  - <X> + the succeeding N words.

Example

- Sentence:
  - ”this was the last of **5 legs** and approaching **the end** of **an 8 hour duty day** and **7 hour hard time flying day**.”
- Phrases and patterns:
  - **5 legs**: last of <X>, <X> and approaching
  - **end**: and approaching <X>, <X> of an
  - **8 hour duty day**: end of <X>, <X> and 7
  - **7 hour hard time flying day**: day and <X>

Parse Tree Based Patterns

- We have used two parsers, the Collins parser and the Minipar parser to parse the ASRS corpus to generate parse trees and discover the dependencies and structural patterns in the text.
- Despite the informal nature of the text with grammatical mistakes, we have been able to parse 95.8% of the text with the Collins parser and 99.62% of the text with the Minipar parser.

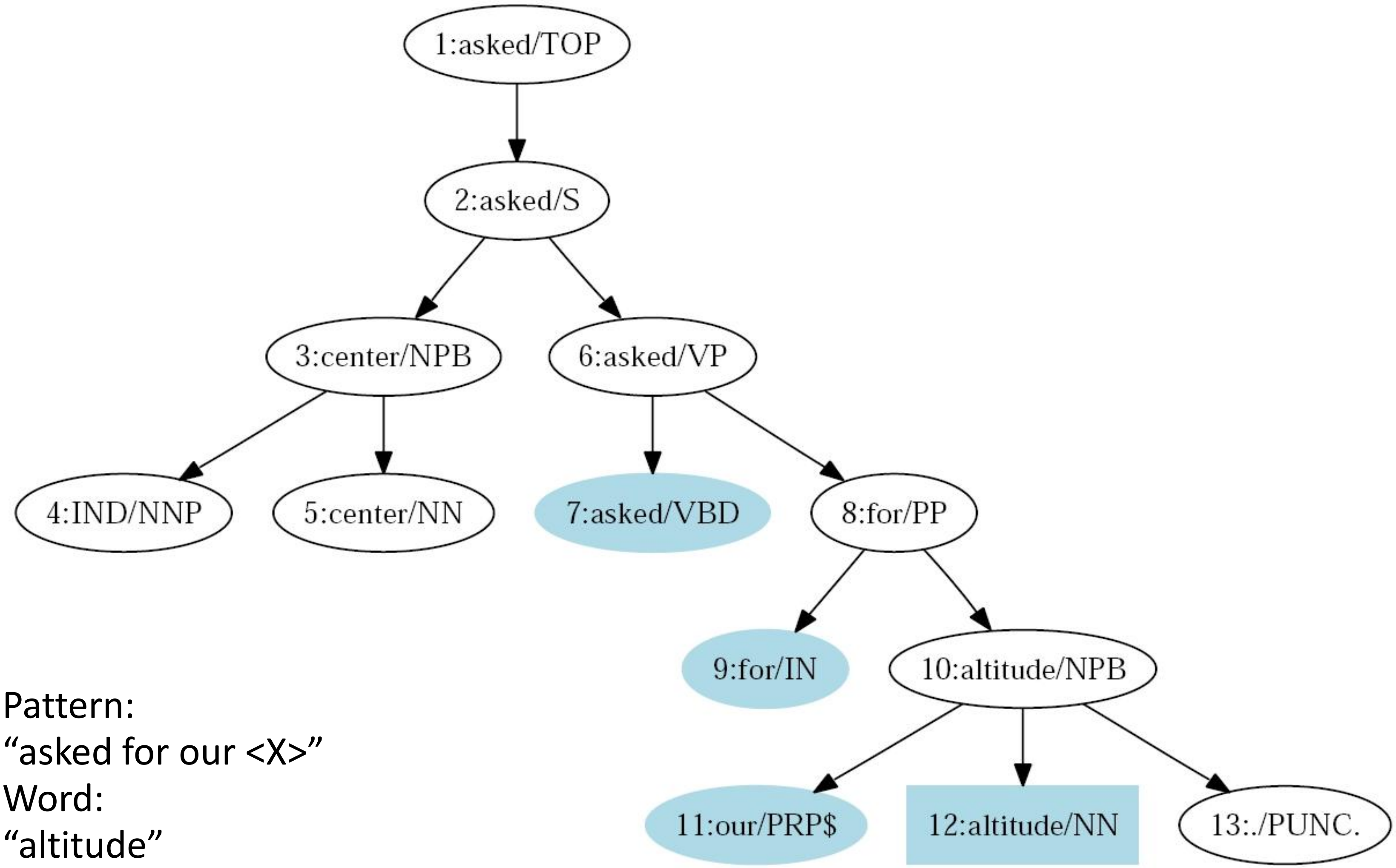
Governing Word

- An interesting information that we can extract from a parse tree is the word that governs a given word.
  - For example, in the sentence ”we were in our climb,” the noun ”climb” is governed by the verb ”were.”
- Nouns are usually governed by verbs, while adjectives are usually governed by nouns or verbs.
- We extract and utilize these information from the parse trees to use as patterns.

Governing Word Patterns

- For each noun and adjective in the ASRS corpus, we find the governing word from the Collins parse tree.
- We use the governing word and the sentence fragment between the governing word and the actual word as the pattern.
- For example, in the sentence ”*IND center asked for our altitude,*” the pattern for the noun ”*altitude*” is ”*asked for our <X>.*”

Example



Pattern:
“asked for our <X>”
Word:
“altitude”

Governing Phrase Patterns

- The governing word concept can be easily extended to phrases.
- We identify the phrases using the CRFChunker tool, and extract governance information using the Collins parser.
- We use the governing phrase and the sentence fragment between the governing phrase and the actual phrase as the pattern.
  - For example, in the sentence ”*a line of thunderstorms were indicated on the weather radar*,” the noun phrase ”*weather radar*” is governed by ”*were indicated,*” giving the pattern “*were indicated on <X>*”

Patterns From MINIPAR

- The MINIPAR parser, in addition to governing word information, identifies the subject and objects in the sentence, and the verb that governs the subjects and objects.
- We used the relationships between subject, verb and object identified by MINIPAR to extract subject-verb-object patterns.
- Also, for adjectives, we use the governing word (noun or verb) and the sentence fragment between them as the pattern.

MINIPAR Patterns

- We use three types of patterns from MINIPAR:
  1. For each subject in the corpus, we find the governing verb from MINIPAR parse tree. Then we use the governing verb and the sentence fragment between the governing verb and the subject as the pattern.
  2. Similarly, for each object in the corpus, the governing verb and the sentence fragment between the governing verb and the subject is used as the pattern.
  3. Also, the governing word and sentence fragment between the governing word and the adjective acts as a pattern.

Example

- Sentence:
  - "Within several seconds **engine** began to make loud **noises** as if a **rod** had broken."
- Words and patterns:
  - **engine**: <X> began
  - **noises**: make loud <X>
  - **rod**: <X> had broken

Merging Lexicons

- The lexicons developed by different patterns do not overlap completely – each lexicon has some words that are not in the other lexicons.
- By merging the different lexicons developed by different patterns, we can compile a bigger lexicon.
- We have combined the lexicons developed by the 7 patterns discussed so far. This has given us a lexicon of 3121 words.

Classifying Reports

- Using the merged lexicon, we followed a simple classification scheme:
  - If a word related to a shaping factor ***S*** appears in the narrative of report ***R***, then label ***R*** with ***S***.
  - This was tested on 7 categories: *Duty Cycle*, *Familiarity*, *Physical Environment*, *Physical Factors*, *Preoccupation*, *Proficiency* and *Resource Deficiency.*
  - The classification was tested on 1000 manually annotated reports.

Classification Performance Of Lexicon From Modified Basilisk

Shaping Factor	TP	FN	TN	FP	Prec	Rec	F
Duty Cycle	13	15	957	15	0.46	0.46	0.46
Familiarity	41	9	382	568	0.07	0.82	0.12
Physical Environment	219	39	578	164	0.57	0.85	0.68
Physical Factors	28	7	828	137	0.17	0.8	0.28
Preoccupation	88	22	644	246	0.26	0.8	0.4
Proficiency	181	67	387	365	0.33	0.73	0.46
Resource Deficiency	338	170	288	204	0.62	0.67	0.64
Overall	**908**	**329**	**4064**	**1699**	**0.35**	**0.73**	**0.47**

Original Basilisk Lexicon

- For comparison, we also ran the original Basilisk framework with the exact same patterns for 10 iterations.
- The classification performance of this lexicon was tested on the same 1000 reports for the 7 shaping factors.

Classification performance of lexicon from original Basilisk

Shaping Factor	TP	FN	TN	FP	Prec	Rec	F
Duty Cycle	28	0	21	951	0.03	1.00	0.06
Familiarity	50	0	19	931	0.05	1.00	0.10
Physical Environment	255	3	47	695	0.27	0.99	0.42
Physical Factors	35	0	20	945	0.04	1.00	0.07
Preoccupation	110	0	16	874	0.11	1.00	0.20
Proficiency	247	1	18	734	0.25	1.00	0.40
Resource Deficiency	500	8	7	485	0.51	0.98	0.67
Overall	**1225**	**12**	**148**	**5615**	**0.18**	**0.99**	**0.30**

Text Classification With Unigram SVM

- We compared our method with supervised text classification method of RBF-kernel SVM with unigrams as features.
- For each of the 7 shaping factors, one SVM classifier was trained on the 233 reports from which the seed words were drawn.
- The SVM classifiers were tested on 1000 reports.

Classification Performance Of SVM Classifier Using Unigram Features

Shaping Factors	TP	FN	TN	FP	Prec	Rec	F
Duty Cycle	6	22	952	20	0.23	0.21	0.22
Familiarity	1	49	950	0	1.00	0.02	0.04
Physical Environment	81	177	677	65	0.55	0.31	0.40
Physical Factors	11	24	943	22	0.33	0.31	0.32
Preoccupation	8	102	859	31	0.21	0.07	0.11
Proficiency	14	234	731	21	0.40	0.06	0.10
Resource Deficiency	277	231	339	153	0.64	0.55	0.59
Overall	**398**	**839**	**5451**	**312**	**0.56**	**0.32**	**0.41**

Topic Model

- The document topic model represents each document as a probability distribution over a number of topics [Blei et al. 2003].
  - Each topic is a probability distribution over a number of words.
  - Application of Latent Dirichlet Allocation (LDA) on the corpus identifies the different topics latent in the corpus.

Topic Model SVM Classifier

- we generated 50 topics from the training set and test set together without considering the actual labels.
- Then we separated the training set from the test set and using the $\gamma^*(w)$ values as features, we trained an SVM classifier on the training set and tested the performance on the test data.

Classification Performance Of SVM Classifier Using 50 Topics

Shaping factor	TP	FN	TN	FP	Prec	Rec	F
Duty Cycle	0	28	972	0	0.00	0.00	0.00
Familiarity	0	50	950	0	0.00	0.00	0.00
Physical Environment	0	258	742	0	0.00	0.00	0.00
Physical Factors	0	35	965	0	0.00	0.00	0.00
Preoccupation	2	108	868	22	0.08	0.02	0.03
Proficiency	0	248	752	0	0.00	0.00	0.00
Resource Deficiency	285	223	338	154	0.65	0.56	0.60
Overall	**287**	**950**	**5587**	**176**	**0.62**	**0.23**	**0.34**

Analysis

- The results of our experiments show that it is feasible to apply NLP techniques to grow a semantic lexicon for the different Shaping Factors.
- The application of part of speech tagging, phrase chunking and parsing to discover sentence structures and dependencies have given us valuable information. Using this information we have been able to grow this semantic lexicon.

Analysis

- The success of our methods show the strength of probabilistic models for POS tagging and shallow parsing, and also the ability of statistical and broad coverage parsers to process informally written text containing grammatical mistakes.
- This shows that it is possible to apply NLP techniques on this corpus to extract useful information for the task of report classification.

Future Works

- Further methods of word ranking and classification:
  - Additional methods of measuring word similarity, for example, the K-L divergence between the pattern probability vectors.
  - Clustering techniques to group similar words together based on extraction pattern features.

Future Works

- New report classification schemes
  - Using the keywords and phrases learned form the text and the seed words, we can explore the following report classification schemes:
    - Probabilistic classification
    - Entailment based classification
    - Neural network based classification

Future Works

- Probabilistic Classification:
  - Compute the probability that the incident N was due to the shaping factor C as follows:
    - P(C|N) = Sum of all P(C|W) * P(W|N) for all W that occurs in the keyword list for shaping factor C and the narrative for incident N.
      - P(W|N) = Probability that word W occurs in the narrative for incident N
      - P(C|W) = Probability that the keyword list for shaping factor C contains W

Future Works

- Entailment based classification
  - Represent the report R of incident N as a Knowledge Base (KB) by converting it into a logical form using predicates
  - For each category C, use logical inference to establish whether the KB entails the predicate CAUSED(N, C)
    - That is, whether we can infer from the KB that C was a shaping factor for N
    - The keyword list for the shaping factors will be needed to generalize the predicates.

Future Works

- Neural Network based classification
  - Represent each report as a vector of features
    - For example, the presence or absence, or probability of appearance, of the keywords related to each shaping factor and also the extraction patterns.
  - Train and use a Neural Network to identify which of the shaping factors contributed to the incident.

References

[Posse 2005] Posse, C., Matzke, B., Anderson, C., Brothers, A., Matzke, M., Ferryman, T., *Extracting information from narratives: an application to aviation safety reports*, Aerospace Conference, 2005 IEEE

[Ferryman 2006] Ferryman, T. A., Posse, C., Rosenthal, L. J., Srivastava, A. N., and Statler, I. C., *What Happened, and Why: Toward an Understanding of Human Error Based on Automated Analyses of Incident Reports—Vol. II*, NASA/TP–2006-213490

References

[Thelen 2002] Thelen, M. and Riloff, E., *A bootstrapping method for learning semantic lexicons using extraction pattern contexts,* Acl-02 Conference on Empirical Methods in Natural Language Processing - Volume 10

[Blei et al. 2003] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 2003.

Resources

- ASRS Data Set:
  - http://asrs.arc.nasa.gov/
- CRFTagger
  - http://crftagger.sourceforge.net
- Collins Parser
  - http://people.csail.mit.edu/mcollins/code.html
- Minipar Parser
  - http://www.cs.ualberta.ca/~lindek/minipar.htm